TITLE:

Log canonical threshold of Vandermonde matrix type singularities and learning theory (Applications of Reproducing Kernels)

AUTHOR(S):

Aoyagi, Miki

CITATION:

Aoyagi, Miki. Log canonical threshold of Vandermonde matrix type singularities and learning theory (Applications of Reproducing Kernels). 数理解析研究所講究録 2008, 1618: 24-40

ISSUE DATE:

2008-12

URL:

http://hdl.handle.net/2433/140196

RIGHT:

京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

数理解析研究所講究録
第1618巻 2008年 24-40

Log canonical threshold of Vandermonde matrix type singularities and learning theory

Miki Aoyagi*

Abstract

In this paper, we consider the log canonical threshold of Vandermonde matrix type singularities over the real field. It has recently been proved that these singularities are essential in learning theory.

1 Introduction

The log canonical threshold $c_Z(Y, f)$ in algebraic geometry is analytically defined by

$$c_Z(Y, f) = \sup\{c : |f|^{-c} \text{ is locally } L^2 \text{ near } Z\},$$

over $\mathbb{C}$ and

$$c_Z(Y, f) = \sup\{c : |f|^{-c} \text{ is locally } L^1 \text{ near } Z\},$$

over $\mathbb{R}$ for a nonzero regular function f on a smooth variety Y, where $Z \subset Y$ is a closed subscheme([16], [19]). It is also known that $c_0(\mathbb{C}^d, f)$ is the largest root of the Bernstein-Sato polynomial $b(s) \in \mathbb{C}[s]$ of f, where $b(s)f^s = Pf^{s+1}$ for a linear differential operator P([8], [9], [15]).

Watanabe proved that the largest pole of a zeta function for a hierarchical learning model gives the main term of the generalization error of the model asymptotically ([24],[25]). The largest pole of $\int_{\text{near } Z} |f|^{2z}\psi(w)dw$ over $\mathbb{C}$ ($\int_{\text{near } Z} |f|^{z}\psi(w)dw$ over $\mathbb{R}$), corresponds to the log canonical threshold $c_Z(Y, f)$, where $\psi(w)$ is a $C^\infty-$ function with a compact support and $\psi(Z) \neq 0$.

The theoretical study of hierarchical learning models has been rapidly developed in recent years. A learning system consists of data, a learning model and a learning algorithm. The purpose of such a system is to estimate an unknown true density function from data distributed by the true density function. The data associated with image or speech recognition, artificial intelligence, the control of a robot, genetic analysis, data mining, time series prediction, and so on, are very complicated and not usually generated by a simple normal distribution, as they are influenced by many factors. Learning models to analyze such data should likewise have complicated structures. Hierarchical learning models such as the layered neural network model, the Boltzmann machine, the reduced rank regression model and the normal mixture model may be known as effective learning

*ARISH, Nihon University, Nihon University Kaikan Daini Bekkan, 12-5, Goban-cho, Chiyoda-ku, Tokyo 102-8251, Japan. email: aoyagi.miki@nihon-u.ac.jp

models. They are, however, non-regular statistical models, which cannot be analyzed using the classic theories of regular statistical models [13], [23], [12], [10]. The theoretical study has therefore been started to construct a mathematical foundation for non-regular statistical models.

The generalization error of a learning model is a difference between a true density function and a predictive density function obtained using distributed training samples. It is one of the most important topic in learning theory. The largest pole of a zeta function for a learning model, which is called a learning coefficient, gives the main term of the generalization error and can be obtained by a desingularization.

In spite of these mathematical foundations, obtaining the largest pole is still difficult for the following reason.

It is known that the desingularization is obtained by using a finite blowing up process [14]. However, desingularization in general is very difficult. Furthermore, most of functions for hierarchical learning models are degenerate with respect to their Newton polyhedrons [11], their singularities are not isolated and they are not simple polynomials, i.e., they have parameters.

We note that there are many classical results for calculating the largest poles of the zeta functions using the desingularization in lower dimension. There have also been many investigations in the case of prehomogeneous spaces. The functions, however, do not occur in prehomogeneous spaces.

Therefore, most of these singularities in learning theory have not been investigated, so far.

Our study is over the real field not the complex field. In algebraic geometry and algebraic analysis, these studies are usually done over an algebraically closed field. We have many differences between the real field and the complex field, for example, log canonical thresholds over the complex field are less than 1, while those over the real field are not necessarily less than 1.

In this paper, we consider the log canonical threshold of Vandermonde matrix type singularities which is the largest pole of zeta functions for the three layered neural network and the normal mixture model, as such models are widely used in many applied fields.

Theorem 1 shows a kind of an orthogonal relation of the log canonical threshold of Vandermonde matrix type singularities. It means that the learning model learns a true distribution independently on each hidden unit in case of three layered neural networks or each peak in case of the normal mixture model (Section 3).

Theorem 2 gives the log canonical thresholds in some condition. Our future purpose is to obtain the log canonical thresholds of Vandermonde matrix type singularities in general.

Recently, the term "algebraic statistics" arises from the study of probabilistic models and techniques for statistical inference using methods from algebra and geometry [22]. Our study may stand for this attitude.

2 Vandermonde matrix type singularities

In this paper, we denote by a^*, b^* constants and denote by a^* if the variable a is in a sufficiently small neighborhood of a^*.

Define the norm of a matrix $C = (c_{ij})$ by $||C|| = \sqrt{\sum_{i,j} |c_{ij}|^2}$. Denote by $\langle C \rangle$ the ideal generated by $\{c_{ij}\}$. Set $\mathbb{N}_{+0} = \mathbb{N} \cup \{0\}$.

Definition 1 *Set* $c_Z(f) = \sup\{c : |f|^{-c}$ *is locally* L^1 *near* $Z\}$ *over* $\mathbb{R}$, *for a nonzero regular function* f *on a neighborhood of* Z, *where* Z *is a closed subscheme.*

Definition 2 *Fix* $Q \in \mathbb{N}$. *Define* $[b_1^*, b_2^*, \cdots, b_N^*]_Q = \gamma_i(0, \cdots, 0, b_i^*, \cdots, b_N^*)$ *if* $b_1^* = \cdots = b_{i-1}^* = 0$, $b_i^* \neq 0$, *and* $\gamma_i = \begin{cases} 1 & \text{if } Q \text{ is odd,} \\ |b_i^*|/b_i^* & \text{if } Q \text{ is even.} \end{cases}$

Definition 3 *Fix* $Q \in \mathbb{N}$ *and* $m \in \mathbb{N}_{+0}$.

$$\text{Let } A = \begin{pmatrix} a_{11} & \cdots & a_{1H} & a_{1,H+1}^* & \cdots & a_{1,H+r}^* \\ a_{21} & \cdots & a_{2H} & a_{2,H+1}^* & \cdots & a_{2,H+r}^* \\ & \vdots & & & \vdots & \\ a_{M1} & \cdots & a_{MH} & a_{M,H+1}^* & \cdots & a_{M,H+r}^* \end{pmatrix}, \ I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N,$$

$$B_I = (\prod_{j=1}^N b_{1j}^{\ell_j}, \prod_{j=1}^N b_{2j}^{\ell_j}, \cdots, \prod_{j=1}^N b_{Hj}^{\ell_j}, \prod_{j=1}^N b_{H+1,j}^{*\ \ell_j}, \cdots, \prod_{j=1}^N b_{H+r,j}^{*\ \ell_j})^t$$

and $B = (B_I)_{\ell_1+\cdots+\ell_N=Qn+m, 0\leq n\leq H+r-1}$ *(t denotes the transpose).*

We call singularities of $||AB||^2 = 0$ *Vandermonde matrix type singularities.*

To simplify, we usually assume that

$$(a_{1,H+j}^*, a_{2,H+j}^*, \cdots, a_{M,H+j}^*)^t \neq 0, (b_{H+j,1}^*, b_{H+j,2}^*, \cdots, b_{H+j,N}^*) \neq 0$$

for $1 \leq j \leq r$ *and*

$$[b_{H+j,1}^*, b_{H+j,2}^*, \cdots, b_{H+j,N}^*]_Q \neq [b_{H+j',1}^*, b_{H+j',2}^*, \cdots, b_{H+j',N}^*]_Q$$

for $j \neq j'$.

From now on, we set A and B as in Definition 3.

Remark 1 *By the ascending chain condition, we have* $\langle AB \rangle = \langle AB' \rangle$ *where* $B' = (B_I)_{\ell_1+\cdots+\ell_N=Qn+m, 0\leq n\leq H'}$ *and* $H' \geq H + r - 1$.

Example 1 *If* $N = 1$, $m = 0$, $Q = 1$ *and* $r = 0$, *we have* $A = \begin{pmatrix} a_{11} & \cdots & a_{1H} \\ a_{21} & \cdots & a_{2H} \\ & \vdots & \\ a_{M1} & \cdots & a_{MH} \end{pmatrix}$ *and*

$$B = \begin{pmatrix} 1 & b_{11} & b_{11}^2 & \cdots & b_{11}^{H-1} \\ 1 & b_{21} & b_{21}^2 & \cdots & b_{21}^{H-1} \\ & & & \vdots & \\ 1 & b_{H1} & b_{H1}^2 & \cdots & b_{H1}^{H-1} \end{pmatrix}.$$

(The matrix B *as above is usually called a Vandermonde matrix.)*

Example 2 *If* $N = 3$, $m = Q = 1$ *and* $r = H = 1$, *we have* $A = \begin{pmatrix} a_{11} & a^*_{12} \\ a_{21} & a^*_{22} \\ \vdots & \vdots \\ a_{M1} & a^*_{M,2} \end{pmatrix}$ *and*

$$B = \begin{pmatrix} b_{11} & b_{11}^2 & b_{12} & b_{12}^2 & b_{13} & b_{13}^2 & b_{11}b_{12} & b_{11}b_{13} & b_{12}b_{13} \\ b^*_{21} & {b^*_{21}}^2 & b_{22} & b_{22}^2 & b_{23} & b_{23}^2 & b^*_{21}b^*_{22} & b^*_{21}b^*_{23} & b^*_{22}b^*_{23} \end{pmatrix}.$$

Theorem 1 *Consider a sufficiently small neighborhood of*

$$w^* = \{a^*_{ki}, b^*_{ij}\}_{1\le k\le M, 1\le i\le H, 1\le j\le N}.$$

Set $(b^{**}_{01}, b^{**}_{02}, \cdots, b^{**}_{0N}) = (0, \ldots, 0)$.
Let each $(b^{**}_{11}, b^{**}_{12}, \cdots, b^{**}_{1N})$, …, $(b^{**}_{r'1}, b^{**}_{r'2}, \cdots, b^{**}_{r'N})$ *be a different real vector in*

$$[b^*_{i1}, b^*_{i2}, \cdots, b^*_{iN}]_Q \neq 0, \text{ for } i = 1, \ldots, H + r :$$

$$\{(b^{**}_{11}, \cdots, b^{**}_{1N}), \ldots, (b^{**}_{r'1}, \cdots, b^{**}_{r'N}) \ ; \ [b^*_{i1}, \cdots, b^*_{iN}]_Q \neq 0, i = 1, \ldots, H + r\}.$$

Then $r' \ge r$ *and set* $(b^{**}_{i1}, \cdots, b^{**}_{iN}) = [b^*_{H+i,1}, \cdots, b^*_{H+i,N}]_Q$, *for* $1 \le i \le r$.
Assume that

$$\left.\begin{matrix} [b^*_{11}, \cdots, b^*_{1N}]_Q \\ \vdots \\ [b^*_{H_0 1}, \cdots, b^*_{H_0 N}]_Q \end{matrix}\right\} = 0,$$

$$\left.\begin{matrix} [b^*_{H_0+1,1}, \cdots, b^*_{H_0+1,N}]_Q \\ \vdots \\ [b^*_{H_0+H_1,1}, \cdots, b^*_{H_0+H_1,N}]_Q \end{matrix}\right\} = (b^{**}_{11}, \cdots, b^{**}_{1N}),$$

$$\left.\begin{matrix} [b^*_{H_0+H_1+1,1}, \cdots, b^*_{H_0+H_1+1,N}]_Q \\ \vdots \\ [b^*_{H_0+H_1+H_2,1}, \cdots, b^*_{H_0+H_1+H_2,N}]_Q \end{matrix}\right\} = (b^{**}_{21}, \cdots, b^{**}_{2N}),$$

$$\vdots$$

$$\left.\begin{matrix} [b^*_{H_0+\cdots+H_{r'-1}+1,1}, \cdots, b^*_{H_0+\cdots+H_{r'-1}+1,N}]_Q \\ \vdots \\ [b^*_{H_0+\cdots+H_{r'-1}+H_{r'},1}, \cdots, b^*_{H_0+\cdots+H_{r'-1}+H_{r'},N}]_Q \end{matrix}\right\} = (b^{**}_{r'1}, \cdots, b^{**}_{r'N}).$$

and $H_0 + \cdots + H_{r'} = H$.
Then we have

$$c_{w^*}(||AB||^2) = \sum_{\alpha=0}^{r'} c_{w^{(\alpha)*}}(||A^{(\alpha)}B^{(\alpha)}||^2),$$

where $w^{(\alpha)*} = \{a^{(\alpha)*}_{ki}, b^{(\alpha)*}_{ij}\} = \{a^*_{k,H_0+\cdots+H_{\alpha-1}+i}, b^{**}_{\alpha j}\}_{1\le k\le M, 1\le i\le H_\alpha, 1\le j\le N}$,

$I = (\ell_1, \ldots, \ell_N) \in \mathbb{N}_{+0}{}^N$,

$$A^{(\alpha)} = \begin{pmatrix} a_{11}^{(\alpha)} & a_{12}^{(\alpha)} & \cdots & a_{1H_\alpha}^{(\alpha)} \\ a_{21}^{(\alpha)} & a_{22}^{(\alpha)} & \cdots & a_{2H_\alpha}^{(\alpha)} \\ & & \vdots & \\ a_{M1}^{(\alpha)} & a_{M2}^{(\alpha)} & \cdots & a_{MH_\alpha}^{(\alpha)} \end{pmatrix}, B_I^{(\alpha)} = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{(\alpha)\ell_j} \\ \prod_{j=1}^N b_{2j}^{(\alpha)\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H_\alpha j}^{(\alpha)\ \ell_j} \end{pmatrix}, \text{ for } \alpha = 0, r+1 \le \alpha \le r',$$

$$A^{(\alpha)} = \begin{pmatrix} a_{11}^{(\alpha)} & a_{12}^{(\alpha)} & \cdots & a_{1H_\alpha}^{(\alpha)} & a_{1,H+\alpha}^{*} \\ a_{21}^{(\alpha)} & a_{22}^{(\alpha)} & \cdots & a_{2H_\alpha}^{(\alpha)} & a_{2,H+\alpha}^{*} \\ & & \vdots & & \\ a_{M1}^{(\alpha)} & a_{M2}^{(\alpha)} & \cdots & a_{MH_\alpha}^{(\alpha)} & a_{M,H+\alpha}^{*} \end{pmatrix}, B_I^{(\alpha)} = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{(\alpha)\ell_j} \\ \prod_{j=1}^N b_{2j}^{(\alpha)\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H_\alpha j}^{(\alpha)\ \ell_j} \\ \prod_{j=1}^N b_{\alpha j}^{**\ell_j} \end{pmatrix}, \text{ for } 1 \le \alpha \le r,$$

$B^{(0)} = (B_I^{(0)})_{\ell_1+\ldots+\ell_N = Qn+m, 0 \le n \le H_0 - 1}$ *and* $B^{(\alpha)} = (B_I^{(\alpha)})_{\ell_1+\ldots+\ell_N = n, 0 \le n \le H_\alpha - 1}$ *for* $1 \le \alpha \le r'$.

(Proof)
Set

$$\begin{cases} (a_{i1}^{(0)}, \ldots, a_{iH_0}^{(0)}) = (a_{i1}, \ldots, a_{iH_0}), \\ (a_{i1}^{(1)}, \ldots, a_{iH_1}^{(1)}) = (a_{i,H_0+1}, \ldots, a_{i,H_0+H_1}), \\ \quad \vdots \\ (a_{i1}^{(r')}, \ldots, a_{iH_{r'}}^{(r')}) = (a_{i,H_0+\cdots+H_{r'-1}+1}, \ldots, a_{i,H_0+\cdots+H_{r'}}), \end{cases} \quad \text{for } 1 \le i \le M, \text{ and}$$

$$\begin{cases} (b_{1j}^{(0)}, \ldots, b_{H_0 j}^{(0)}) = (b_{1j}, \ldots, b_{H_0 j}), \\ (b_{1j}^{(1)}, \ldots, b_{H_1 j}^{(1)}) = (b_{H_0+1,j}, \ldots, b_{H_0+H_1,j}), \\ \quad \vdots \\ (b_{1j}^{(r')}, \ldots, b_{H_{r'} j}^{(r')}) = (b_{H_0+\cdots+H_{r'-1}+1,j}, \ldots, b_{H_0+\cdots+H_{r'},j}), \end{cases} \quad \text{for } 1 \le j \le N.$$

For $\gamma_i(b_{i1}^{(\alpha)}, \cdots, b_{iN}^{(\alpha)}) = [b_{i1}^{(\alpha)}, \cdots, b_{iN}^{(\alpha)}]_Q$, we again set $a_{ki}^{(\alpha)}$ by $a_{ki}^{(\alpha)}/(\gamma_i)^m$ and $b_{ij}^{(\alpha)}$ by $b_{ij}^{(\alpha)}\gamma_i$, $1 \le j \le N$ and $1 \le k \le M$.

Main parts of the proof is appeared in Appendix. By applying Lemma 4 in Appendix we have this theorem.

Q.E.D.

Usually, r corresponds to the number of elements of a true distribution. This theorem shows that the Bayesian learning coefficient related with such singularities is the sum of each for the small model with respect to each element of a true distribution (cf. Section 3).

Theorem 2 *We use the same notations as in Theorem 1. If* $N = 1$, *we have*

$$\begin{aligned} c_{w^*}(||AB||^2) &= \frac{MQk_0(k_0+1) + 2H_0}{4(m+k_0Q)} \\ &+ \frac{Mr'}{2} + \sum_{\alpha=1}^{r} \frac{Mk_\alpha(k_\alpha+1) + 2H_\alpha}{4(1+k_\alpha)} + \sum_{\alpha=r+1}^{r'} \frac{Mk_\alpha(k_\alpha+1) + 2(H_\alpha - 1)}{4(1+k_\alpha)}, \end{aligned}$$

where

$$k_0 = \max\{i \in \mathbb{Z}; 2H_0 \geq M(i(i-1)Q + 2mi)\},$$
$$k_\alpha = \max\{i \in \mathbb{Z}; 2H_\alpha \geq M(i^2 + i)\}, \ \textit{for } 1 \leq \alpha \leq r,$$
$$k'_\alpha = \max\{i \in \mathbb{Z}; 2(H_\alpha - 1) \geq M(i^2 + i)\}, \ \textit{for } r+1 \leq \alpha \leq r'.$$

For the proof of Theorem 2, we use a similar method in [6], [4], where we used recursive blowing ups and toric resolution.

The key point is that $c_0(||A^{(0)}B^{(0)}||^2) = c_0(||A^{(0)}B'||^2)$ for $N = 1$, where

$$B' = \begin{pmatrix} b_{11}^m & 0 & 0 & \cdots & 0 \\ 0 & b_{21}^m(b_{21}^Q - b_{11}^Q) & 0 & \cdots & 0 \\ 0 & 0 & b_{31}^m(b_{31}^Q - b_{11}^Q)(b_{31}^Q - b_{21}^Q) & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & b_{H1}^m(b_{H1}^Q - b_{11}^Q)\cdots(b_{H1}^Q - b_{H-1,1}^Q) \end{pmatrix},$$

and $|b_{H1}| < |b_{H-1,1}| < \cdots < |b_{21}| < |b_{11}|$.

Recently, we have the explicit values $c_{w^*}(||AB||^2)$ for general natural numbers N and M but for $H \leq 2$ [5].

The following is also an important learning model, which is called reduced rank regression. The model corresponds to the three-layer neural network with linear hidden units.

Theorem 3 ([7]) *Let* $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} & a^*_{1,H+1} & \cdots & a^*_{1,H+r} \\ a_{21} & a_{22} & \cdots & a_{2H} & a^*_{2,H+1} & \cdots & a^*_{2,H+r} \\ & & \vdots & & \vdots & & \\ a_{M1} & a_{M2} & \cdots & a_{MH} & a^*_{M,H+1} & \cdots & a^*_{M,H+r} \end{pmatrix}$ *and*

$$B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1N} \\ b_{21} & b_{22} & \cdots & b_{2N} \\ & & \vdots & \\ b_{H1} & b_{H2} & \cdots & b_{HN} \\ b^*_{H+1,1} & b^*_{H+1,2} & \cdots & b^*_{H+1,N} \\ & & \vdots & \\ b^*_{H+r,1} & b^*_{H+r,2} & \cdots & b^*_{H+r,N} \end{pmatrix}.$$

Let r *be the rank of* $\begin{pmatrix} a^*_{1,H+1} & \cdots & a^*_{1,H+r} \\ a^*_{2,H+1} & \cdots & a^*_{2,H+r} \\ & \vdots & \\ a^*_{M,H+1} & \cdots & a^*_{M,H+r} \end{pmatrix} \begin{pmatrix} b^*_{H+1,1} & b^*_{H+1,2} & \cdots & b^*_{H+1,N} \\ & & \vdots & \\ b^*_{H+r,1} & b^*_{H+r,2} & \cdots & b^*_{H+r,N} \end{pmatrix}$.

Then the log canonical threshold of $||AB||^2$ *at* $Z = \{||AB||^2 = 0\}$ *is*

$$\max\{-\frac{(N+M)r - r^2 + s(N-r) + (M-r-s)(H-r-s)}{2} \mid 0 \leq s \leq \min\{M+r, H+r\}\}.$$

That is,

Case 1 *Let* $N+r \leq M+H$, $M+r \leq N+H$ *and* $H+r \leq M+N$.
(a) If $M+H+N+r$ *is even, then*

$$c_Z(||AB||^2) = \frac{-(H+r)^2 - M^2 - N^2 + 2(H+r)M + 2(H+r)N + 2MN}{8}.$$

(b) If $M+H+N+r$ *is odd, then*

$$c_Z(||AB||^2) = \frac{-(H+r)^2 - M^2 - N^2 + 2(H+r)M + 2(H+r)N + 2MN + 1}{8}.$$

Case 2 *Let* $M+H < N+r$. *Then* $c_Z(||AB||^2) = \frac{HM - Hr + Nr}{2}$.

Case 3 *Let* $N+H < M+r$. *Then* $c_Z(||AB||^2) = \frac{HN - Hr + Mr}{2}$.

Case 4 *Let* $M+N < H+r$. *Then* $c_Z(||AB||^2) = \frac{MN}{2}$.

3 Learning theorem

In this section, we overview the stochastic complexity and the generalization error in Bayesian estimation.

Let $q(x)$ be a true probability density function and $(x)^n := \{x_i\}_{i=1}^n$ be n training independent and identical samples from $q(x)$. Consider a learning model which is written by a probability form $p(x|w)$, where w is a parameter. The purpose of the learning system is to estimate $q(x)$ from $(x)^n$ by using $p(x|w)$.

Let $p(w|(x)^n)$ be the *a posteriori* probability density function:

$$p(w|(x)^n) = \frac{1}{Z_n}\psi(w)\prod_{i=1}^{n} p(x_i|w),$$

where $\psi(w)$ is an *a priori* probability density function on the parameter set W and

$$Z_n = \int_W \psi(w)\prod_{i=1}^{n} p(x_i|w)\mathrm{d}w.$$

So the average inference $p(x|(x)^n)$ of the Bayesian density function is given by

$$p(x|(x)^n) = \int p(x|w)p(w|(x)^n)\mathrm{d}w,$$

which is the predictive density function.

Set

$$K(q||p) = \int q(x)\log\frac{q(x)}{p(x|(x)^n)}\mathrm{d}x.$$

This is always a positive value and satisfies $K(q||p) = 0$ if and only if $q(x) = p(x|(x)^n)$.

The generalization error $G(n)$ is its expectation value E_n over n training samples:

$$G(n) = E_n\{\int q(x) \log \frac{q(x)}{p(x|(x)^n)} \mathrm{d}x\}.$$

Let

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(x)}{p(x_i|w)}.$$

The average stochastic complexity or the free energy is defined by

$$F(n) = -E_n\{\log \int \exp(-nK_n(w))\psi(w)\mathrm{d}w\}.$$

Then we have $G(n) = F(n+1) - F(n)$ for an arbitrary natural number n ([17], [2], [3]). $F(n)$ is known as the Bayesian criterion in Bayesian model selection [21], stochastic complexity in universal coding [20], [28], Akaike's Bayesian criterion in optimization of hyperparameters [1] and evidence in neural network learning [18].

It has recently been proved that the largest pole of a zeta function gives the generalization error of hierarchical learning models asymptotically [24],[25]. We assume that the true density distribution $q(x)$ is included in the learning model, i.e., $q(x) = p(x|w_t^*)$ for $w_t^* \in W$, where W is the parameter space.

Theorem 4 (Watanabe[24, 25]) *Define the zeta function $J(z)$ of a complex variable z for the learning model by*

$$J(z) = \int K(w)^z \psi(w)\mathrm{d}w,$$

where $K(w)$ is the Kullback function:

$$K(w) = \int p(x|w_t^*) \log \frac{p(x|w_t^*)}{p(x|w)} \mathrm{d}x.$$

Then, for the largest pole $-\lambda$ of $J(z)$ and its order θ, we have

$$F(n) = \lambda \log n - (\theta - 1) \log \log n + O(1), \tag{1}$$

where $O(1)$ is a bounded function of n, and if $G(n)$ has an asymptotic expansion,

$$G(n) \cong \frac{\lambda}{n} - \frac{\theta - 1}{n \log n} \text{ as } n \to \infty. \tag{2}$$

To prove the above theorem, Watanabe used the function

$$v(t) = \int \delta(t - K(w))\varphi(w)\mathrm{d}w = \frac{\partial}{\partial t} \int_{K(w)<t} \varphi(w)\mathrm{d}w,$$

which satisfies $\int v(t)f(t)dt = \int f(K(w))\psi(w)dw$ for any analytic function $f(t)$. The Laplace transform of $v(t)$ is

$$Z(n) = \int \exp(-nK(w))\varphi(w)dw,$$

and the Mellin transform of $v(t)$ is

$$\zeta(z) = \int K(w)^z \varphi(w) dw = \int t^z v(t) dt.$$

The key point of the proof is that by using poles of $\zeta(z)$ and the inverse Mellin transform of $\zeta(z)$, he obtained the asymptotic expansion of $v(t)$, and then the asymptotic expansion of $Z(n)$. The analysis of the difference between $-\log Z(n)$ and $F(n)$ completes the proof.

In learning theory, λ is, therefore, an essential value, which corresponds to the log canonical threshold of $K(w)$.

The log canonical thresholds of Vandermonde matrix type singularities are equal to λ of the following two hierarchical learning models.

(a) The three layered neural network with N input units, H hidden units and M output units which is trained for estimating the true distribution with r hidden units:

Denote an input value by $x = (x_j) \in \mathbb{R}^N$ with a probability density function $q(x)$ which has a compact support $\tilde{W}$. Then an output value $y = (y_k) \in \mathbb{R}^M$ of the three layered neural network is given by $y_k = f_k(x, w) + (\text{noise})$, where $w = \{a_{ki}, b_{ij}; 1 \leq k \leq M, 1 \leq i \leq H, 1 \leq j \leq N\}$ and

$$f_k(x, w) = \sum_{i=1}^{H} a_{ki} \tanh(\sum_{j=1}^{N} b_{ij} x_j).$$

Consider a statistical model

$$p(y|x, w) = \frac{1}{(2\pi)^{M/2}} \exp(-\frac{1}{2}||y - f(x, w)||^2).$$

Assume that the true distribution

$$p(y|x, w_t^*) = \frac{1}{(2\pi)^{M/2}} \exp(-\frac{1}{2}||y - f(x, w_t^*)||^2),$$

is included in the learning model, where $w_t^* = \{a_{ki}^*, b_{ij}^*; 1 \leq k \leq M, H+1 \leq i \leq H+r, 1 \leq j \leq N\}$ and $f_k(x, w_t^*) = \sum_{i=H+1}^{H+r} (-a_{ki}^*) \tanh(\sum_{j=1}^{N} b_{ij}^* x_j)$. Suppose that an *a priori* probability density function $\psi(w)$ is a $C^\infty-$ function with a compact support W where $\psi(w_t^*) > 0$. Then the model has the zeta function $\int_W ||AB||^{2z} dw$ with $Q = 2$ and $m = 1$, where A and B are defined in Definition 3.

(b) The normal mixture model with H peaks which is trained for estimating the true distribution with r peaks [27]:

Consider a normal mixture model

$$p(x|w) = \frac{1}{(2\pi)^{N/2}} \sum_{i=1}^{H} a_{1i} \exp(-\frac{\sum_{j=1}^{N} (x_j - b_{ij})^2}{2}),$$

where $w = \{a_{1i}, b_{ij}; 1 \leq i \leq H, 1 \leq j \leq N\}$ and $\sum_{i=1}^{H} a_{1i} = 1$. Set the true distribution by

$$p(x|w_t^*) = \frac{1}{(2\pi)^{N/2}} \sum_{i=H+1}^{H+r} (-a_{1i}^*) \exp(-\frac{\sum_{j=1}^{N} (x_j - b_{ij}^*)^2}{2}),$$

where $w_t^* = \{a_{1i}^*, b_{ij}^*; H+1 \leq i \leq H+r, 1 \leq j \leq N\}$ and $\sum_{i=H+1}^{H+r} a_{1i}^* = -1$. Suppose that an *a priori* probability density function $\psi(w)$ is a $C^\infty-$ function with a compact support W where $\psi(w_t^*) > 0$.

Then the model has the zeta function $\int_W ||AB||^{2z} dw$ with $Q = 1$, $M = 1$ and $m = 1$, where A and B are defined in Definition 3.

(a) and (b) as above show that λ in Theorem 4 for three layered neural networks and for normal mixture models are obtained by the same type of singularities, i.e., Vandermonde matrix type singularities. The paper [29], moreover, shows that λ for mixtures of binomial distributions is also obtained by Vandermonde matrix type singularities. These facts seem to imply that Vandermonde matrix type singularities are essential for learning theory.

Appendix

Lemma 1 *Let U be a neighborhood of $w^* \in \mathbb{R}^d$. Let $\mathcal{I}$ be the ideal generated by $f_1, \ldots, f_n$ which are analytic functions defined on U. If $g_1, \ldots, g_m \in I$, then $c_{w^*}(f_1^2 + \cdots + f_n^2)$ is greater than $c_{w^*}(g_1^2 + \cdots + g_m^2)$. In particular, if $g_1, \ldots, g_m$ generate the ideal $\mathcal{I}$ then*

$$c_{w^*}(f_1^2 + \cdots + f_n^2) = c_{w^*}(g_1^2 + \cdots + g_m^2).$$

(Proof)

The fact $g_1^2 + \cdots + g_m^2 \leq P(f_1^2 + \cdots + f_n^2)$ for $P >> 1$ yields this lemma.

Q.E.D.

Lemma 2 *Let* $B' = \begin{pmatrix} b_1^m & b_1^{Q+m} & \cdots & b_1^{Q(H-1)+m} \\ \vdots & & & \vdots \\ b_H^m & b_H^{Q+m} & \cdots & b_H^{Q(H-1)+m} \end{pmatrix}$ *and* $\mathbf{b}'_j = \begin{pmatrix} b_1^{Q(j-1)+m} \\ \vdots \\ b_H^{Q(j-1)+m} \end{pmatrix}$.

Consider a sufficiently small neighborhood of $\{b_i^*\}_{1 \leq i \leq H}$.

Let $b_i^* = \gamma_i |b_i^*|$.

Set $\mathbf{b}''_{ij} = \begin{cases} \gamma_i^m \prod_{|b_k^*|=|b_i^*|, 1 \leq k \leq j-1} (b_k/\gamma_k - b_i/\gamma_i), & \text{if } b_i^* \neq 0, \\ b_i^m \prod_{b_k^*=0, 1 \leq k \leq j-1} (b_k^Q - b_i^Q), & \text{if } b_i^* = 0, \end{cases}$ *for $1 \leq j \leq i$ and* $\mathbf{b}''_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{b}''_{jj} \\ \vdots \\ \mathbf{b}''_{Hj} \end{pmatrix}$, *for $1 \leq j \leq H$.*

Then there exists a regular matrix R such that $B'R = (\ \mathbf{b}''_1, \mathbf{b}''_2, \ldots, \mathbf{b}''_H\ )$.

(Proof) We only need to prove that the vector space generated by $\mathbf{b}''_1, \mathbf{b}''_2, \ldots, \mathbf{b}''_H$ is equal to that generated by $\mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_H$.

Some computation shows that the vector space generated by

$$\begin{pmatrix} b_1^m \\ \vdots \\ b_H^m \end{pmatrix}, \begin{pmatrix} 0 \\ b_2^m(b_1^Q - b_2^Q) \\ \vdots \\ b_H^m(b_1^Q - b_H^Q) \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ b_3^m(b_1^Q - b_3^Q)(b_2^Q - b_3^Q) \\ \vdots \\ b_H^m(b_1^Q - b_H^Q)(b_2^Q - b_H^Q) \end{pmatrix}, \cdots, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b_1^m(b_1^Q - b_H^Q)\cdots(b_{H-1}^Q - b_H^Q) \end{pmatrix}$$

is equal to that generated by $\mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_H$.

Therefore, we may set

$$\mathbf{b}'_1 = \begin{pmatrix} b_1^m \\ \vdots \\ b_H^m \end{pmatrix}, \mathbf{b}'_2 = \begin{pmatrix} 0 \\ b_2^m(b_1^Q - b_2^Q) \\ \vdots \\ b_H^m(b_1^Q - b_H^Q) \end{pmatrix}, \cdots, \mathbf{b}'_H = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b_H^m(b_1^Q - b_H^Q)\cdots(b_{H-1}^Q - b_H^Q) \end{pmatrix}.$$

We use an induction.

From now on, denote by $\langle \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_H \rangle$ the vector space generated by vectors $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_H$.

It is easy to check that $\langle \mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_H \rangle = \langle \mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_{H-1}, \mathbf{b}''_H \rangle$.

Let $g_{j,j}(x), g_{j+1,j}(x), \ldots, g_{H,j}(x)$ be polynomials of $x, b_{j-1}, \ldots, b_1$ such that $g_{j',j}(x\gamma_{j'}) = g_{j'',j}(x\gamma_{j''})$ if $|b^*_{j'}| = |b^*_{j''}| \neq 0$ and $g_{j',j}(x) - g_{j'',j}(x')$ can be devided by $x^Q - x'^Q$ if $b^*_{j'} = b^*_{j''} = 0$.

Assume that $\begin{pmatrix} 0 \\ \vdots \\ 0 \\ g_{j,j}(b_j)\mathbf{b}''_{jj} \\ \vdots \\ g_{H,j}(b_H)\mathbf{b}''_{Hj} \end{pmatrix}$ is an element of $\langle \mathbf{b}''_j, \ldots, \mathbf{b}''_H \rangle$ and that

$$\langle \mathbf{b}'_1, \cdots, \mathbf{b}'_H \rangle = \langle \mathbf{b}'_1, \cdots, \mathbf{b}'_{j-1}, \mathbf{b}''_j, \cdots, \mathbf{b}''_H \rangle.$$

Since

$$\mathbf{b}'_{j-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b_{j-1}^m(b_1^Q - b_{j-1}^Q)\cdots(b_{j-2}^Q - b_{j-1}^Q) \\ \vdots \\ b_H^m(b_1^Q - b_H^Q)\cdots(b_{j-2}^Q - b_H^Q) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ g_{j-1,j-1}(b_{j-1})\mathbf{b}''_{j-1,j-1} \\ \vdots \\ g_{H,j-1}(b_H)\mathbf{b}''_{H,j-1} \end{pmatrix},$$

where

$$g_{j-1,j-1}(b_{j-1}) \neq 0, \ldots, g_{H,j-1}(b_H) \neq 0,$$

$g_{j',j-1}(\gamma_{j'}x) = g_{j'',j-1}(\gamma_{j''}x)$ if $|b^*_{j'}| = |b^*_{j''}| \neq 0$ and $g_{j',j-1}(x) - g_{j'',j-1}(x')$ can be divided

by $x'^Q - x^Q$ if $b^*_{j'} = b^*_{j''} = 0$, we have

$$\mathbf{b}'_{j-1} = \mathbf{b}''_{j-1} g_{j-1,j-1}(b_{j-1}) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ (g_{j,j-1}(b_j) - g_{j-1,j-1}(b_{j-1}))\mathbf{b}''_{j,j-1} \\ \vdots \\ (g_{H,j-1}(b_H) - g_{j-1,j-1}(b_{j-1}))\mathbf{b}''_{H,j-1} \end{pmatrix}$$

$$= \mathbf{b}''_{j-1} g_{j-1,j-1}(b_{j-1}) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ g_{j,j}(b_j)\mathbf{b}''_{j,j} \\ \vdots \\ g_{H,j}(b_H)\mathbf{b}''_{H,j} \end{pmatrix},$$

where $\begin{cases} g_{k,j}(b_k) = g_{k,j-1}(b_k) - g_{j-1,j-1}(b_{j-1}), & \text{if } |b^*_k| \neq |b^*_{j-1}|, \\ g_{k,j}(b_k) = (g_{k,j-1}(b_k) - g_{j-1,j-1}(b_{j-1}))/(b_{j-1}/\gamma_{j-1} - b_k/\gamma_k), & \text{if } |b^*_k| = |b^*_{j-1}| \neq 0, \\ g_{k,j}(b_k) = (g_{k,j-1}(b_k) - g_{j-1,j-1}(b_{j-1}))/(b^Q_{j-1} - b^Q_k) & \text{if } b^*_k = b^*_{j-1} = 0. \end{cases}$

By the inductive assumption, $\begin{pmatrix} 0 \\ \vdots \\ 0 \\ g_{j,j}(b_j)\mathbf{b}''_{j,j} \\ \vdots \\ g_{H,j}(b_H)\mathbf{b}''_{H,j} \end{pmatrix}$ is an element of the vector space generated by $\mathbf{b}''_j, \cdots, \mathbf{b}''_H$.

Therefore, $\langle \mathbf{b}'_1, \cdots, \mathbf{b}'_H \rangle = \langle \mathbf{b}'_1, \cdots, \mathbf{b}'_{j-1}, \mathbf{b}''_j, \cdots, \mathbf{b}''_H \rangle = \langle \mathbf{b}'_1, \cdots, \mathbf{b}'_{j-2}, \mathbf{b}''_{j-1}, \mathbf{b}''_j, \cdots, \mathbf{b}''_H \rangle$.

Q.E.D.

Lemma 3 *Let* $B' = \begin{pmatrix} b_1^m & b_1^{Q+m} & \cdots & b_1^{Q(H-1)+m} \\ \vdots & & & \vdots \\ b_H^m & b_H^{Q+m} & \cdots & b_H^{Q(H-1)+m} \end{pmatrix}$ *and* $\mathbf{b}'_j = \begin{pmatrix} b_1^{Q(j-1)+m} \\ \vdots \\ b_H^{Q(j-1)+m} \end{pmatrix}$.

Consider a sufficiently small neighborhood of $\{b^*_i\}_{1 \leq i \leq H}$.

Let $b^*_i = \gamma_i |b^*_i|$.

Let each $|b^{**}_1|, \ldots, |b^{**}_r|$ *be a different real number in* $\{|b^*_i| \ ; \ |b^*_i| \neq 0\}$:

$$\{|b^{**}_1|, \ldots, |b^{**}_r| \ ; \ |b^{**}_i| \neq |b^{**}_j|, i \neq j\} = \{|b^*_i| \ ; \ |b^*_i| \neq 0\}.$$

Also set $b^{**}_0 = 0$.

Assume that $b^*_1 = \cdots = b^*_{H_0} = b^{**}_0$, $|b^*_{H_0+1}| = \cdots = |b^*_{H_0+H_1}| = |b^{**}_1|$, $\ldots$, $|b^*_{H_0+\cdots+H_{r-1}+1}| = \cdots = |b^*_{H_0+\cdots+H_r}| = |b^{**}_r|$.

Set

$$(b^{(0)}_1, \ldots, b^{(0)}_{H_0}) = (b_1, \ldots, b_{H_0}),$$
$$(b^{(1)}_1, \ldots, b^{(1)}_{H_1}) = (b_{H_0+1}, \ldots, b_{H_0+H_1}),$$
$$\vdots$$
$$(b^{(r)}_1, \ldots, b^{(r)}_{H_r}) = (b_{H_0+\cdots+H_{r-1}+1}, \ldots, b_{H_0+\cdots+H_r}).$$

Let $b_i^{(\alpha)^*} = \gamma_i^{(\alpha)} |b_i^{(\alpha)^*}|$.

Then there exists a regular matrix R *such that* $B'R = \begin{pmatrix} B^{(0)} & 0 & 0 & \cdots & 0 \\ 0 & B^{(1)} & 0 & \cdots & 0 \\ & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & B^{(r)} \end{pmatrix}$,

where $B^{(0)} = \begin{pmatrix} b_1^{(0)^m} & b_1^{(0)^{Q+m}} & \cdots & b_1^{(0)^{Q(H_0-1)+m}} \\ \vdots & \vdots & & \vdots \\ b_{H_0}^{(0)^m} & b_{H_0}^{(0)^{Q+m}} & \cdots & b_{H_0}^{(0)^{Q(H_0-1)+m}} \end{pmatrix}$ *and*

$$B^{(\alpha)} = \begin{pmatrix} \gamma_1^{(\alpha)^m} & \gamma_1^{(\alpha)^m} b_1^{(\alpha)}/\gamma_1^{(\alpha)} & \gamma_1^{(\alpha)^m} (b_1^{(\alpha)}/\gamma_1^{(\alpha)})^2 & \cdots & \gamma_1^{(\alpha)^m} (b_1^{(\alpha)}/\gamma_1^{(\alpha)})^{H_\alpha - 1} \\ \vdots & \vdots & & \vdots & \\ \gamma_{H_\alpha}^{(\alpha)^m} & \gamma_{H_\alpha}^{(\alpha)^m} b_{H_\alpha}^{(\alpha)}/\gamma_{H_\alpha}^{(\alpha)} & \gamma_{H_\alpha}^{(\alpha)^m} (b_{H_\alpha}^{(\alpha)}/\gamma_{H_\alpha}^{(\alpha)})^2 & \cdots & \gamma_{H_\alpha}^{(\alpha)^m} (b_{H_\alpha}^{(\alpha)}/\gamma_{H_\alpha}^{(\alpha)})^{H_\alpha - 1} \end{pmatrix}$$

for $1 \le \alpha \le r$.

(Proof)

Set $\mathbf{b}''^{(0)}_1 = \begin{pmatrix} b_1^{(0)^m} \\ b_2^{(0)^m} \\ \vdots \\ b_{H_0}^{(0)^m} \end{pmatrix}$ and $\mathbf{b}''^{(0)}_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b_j^{(0)^m} \prod_{1 \le k \le j-1} (b_k^{(0)^Q} - b_j^{(0)^Q}) \\ \vdots \\ b_{H_0}^{(0)^m} \prod_{1 \le k \le j-1} (b_k^{(0)^Q} - b_{H_0}^{(0)^Q}) \end{pmatrix}$ for $j \ge 2$.

Also set , $\mathbf{b}''^{(\alpha)}_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \gamma_j^{(\alpha)^m} \prod_{1 \le k \le j-1} (b_k^{(\alpha)}/\gamma_k^{(\alpha)} - b_j^{(\alpha)}/\gamma_j^{(\alpha)}) \\ \vdots \\ \gamma_{H_\alpha}^{(\alpha)^m} \prod_{1 \le k \le j-1} (b_k^{(\alpha)}/\gamma_k^{(\alpha)} - b_H^{(\alpha)}/\gamma_H^{(\alpha)}) \end{pmatrix}$ for $1 \le \alpha \le r$, $2 \le j \le i$.

Then, by Lemma 2, there exists a regular matrix R such that

$$B'R = \begin{pmatrix} \mathbf{b}''^{(0)}_1 & \mathbf{b}''^{(0)}_2 & \cdots & \mathbf{b}''^{(0)}_{H_0} & 0 & & \cdots & & & 0 \\ \mathbf{b}''^{(1)}_1 & \mathbf{b}''^{(1)}_1 & \cdots & \mathbf{b}''^{(1)}_1 & \mathbf{b}''^{(1)}_1 & \mathbf{b}''^{(1)}_2 & \cdots & \mathbf{b}''^{(1)}_{H_1} & 0 & \cdots & 0 \\ & & \vdots & & & & \vdots & & & & \\ \mathbf{b}''^{(r)}_1 & \mathbf{b}''^{(r)}_1 & \cdots & \mathbf{b}''^{(r)}_1 & \mathbf{b}''^{(r)}_1 & \mathbf{b}''^{(r)}_1 & \cdots & \mathbf{b}''^{(r)}_1 & \cdots & \mathbf{b}''^{(r)}_1 & \cdots & \mathbf{b}''^{(r)}_{H_r} \end{pmatrix}.$$

Therefore, we have

$$B'RR' = \begin{pmatrix} \mathbf{b}''^{(0)}_1 & \mathbf{b}''^{(0)}_2 & \cdots & \mathbf{b}''^{(0)}_{H_0} & 0 & & \cdots & & & 0 \\ 0 & 0 & \cdots & 0 & \mathbf{b}''^{(1)}_1 & \mathbf{b}''^{(1)}_2 & \cdots & \mathbf{b}''^{(1)}_{H_1} & 0 & \cdots & 0 \\ & & \vdots & & & & \vdots & & & & \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & \mathbf{b}''^{(r)}_1 & \cdots & \mathbf{b}''^{(r)}_{H_r} \end{pmatrix},$$

for some regular matrix R'.

By applying Lemma 2 to $B^{(\alpha)}$, we have the proof.

Q.E.D.

Lemma 4 *Let* $B_I = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{\ell_j} \\ \prod_{j=1}^N b_{2j}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{Hj}^{\ell_j} \end{pmatrix}$
and $B = (B_I)_{\ell_1+\ldots+\ell_N=Q(n-1)+m,n\in\mathbb{N}}$.

Consider a sufficiently small neighborhood of $\{b_{ij}^*\}_{1\le i\le H,1\le j\le N}$.

Let each $(b_{11}^{**}, b_{12}^{**}, \cdots, b_{1N}^{**})$, ..., $(b_{r1}^{**}, b_{r2}^{**}, \cdots, b_{rN}^{**})$ *be a different real vector in*

$$[b_{i1}^*, b_{i2}^*, \cdots, b_{iN}^*]_Q \neq 0, i = 1, \ldots, H + r :$$

$$\{(b_{11}^{**}, \cdots, b_{1N}^{**}), \ldots, (b_{r1}^{**}, \cdots, b_{rN}^{**})\} = \{[b_{i1}^*, \cdots, b_{iN}^*]_Q \neq 0 \ ; \ i = 1, \ldots, H\}.$$

Set $(b_{01}^{**}, b_{02}^{**}, \cdots, b_{0N}^{**}) = (0, \ldots, 0)$.

Assume that

$$[b_{11}^*, \cdots, b_{1N}^*]_Q = \cdots = [b_{H_01}^*, \cdots, b_{H_0N}^*]_Q = (b_{01}^{**}, \cdots, b_{0N}^{**}),$$
$$[b_{H_0+1,1}^*, \cdots, b_{H_0+1,N}^*]_Q = \cdots = [b_{H_0+H_1,1}^*, \cdots, b_{H_0+H_1,N}^*]_Q = (b_{11}^{**}, \cdots, b_{1N}^{**}),$$
$$\ldots,$$
$$[b_{H_0+\cdots+H_{r-1}+1,1}^*, \cdots, b_{H_0+\cdots+H_{r-1}+1,N}^*]_Q = \cdots = [b_{H_0+\cdots+H_r,1}^*, \cdots, b_{H_0+\cdots+H_r,N}^*]_Q = (b_{r1}^{**}, \cdots, b_{rN}^{**}).$$

Set

$$(b_{1j}^{(0)}, \ldots, b_{H_0j}^{(0)}) = (b_{1j}, \ldots, b_{H_0j}),$$
$$(b_{1j}^{(1)}, \ldots, b_{H_1j}^{(1)}) = (b_{H_0+1,j}, \ldots, b_{H_0+H_1,j}),$$
$$\vdots$$
$$(b_{1j}^{(r)}, \ldots, b_{H_rj}^{(r)}) = (b_{H_0+\cdots+H_{r-1}+1,j}, \ldots, b_{H_0+\cdots+H_r,j}),$$

for $1 \le j \le N$.

Let $I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N$, $B_I^{(\alpha)} = \begin{pmatrix} {\gamma_1^{(\alpha)}}^{m-|I|} \prod_{j=1}^N {b_{1j}^{(\alpha)}}^{\ell_j} \\ {\gamma_2^{(\alpha)}}^{m-|I|} \prod_{j=1}^N {b_{2j}^{(\alpha)}}^{\ell_j} \\ \vdots \\ {\gamma_{H_\alpha}^{(\alpha)}}^{m-|I|} \prod_{j=1}^N {b_{H_\alpha j}^{(\alpha)}}^{\ell_j} \end{pmatrix}$
and $B^{(0)} = (B_I^{(0)})_{\ell_1+\ldots+\ell_N=m+Q(n-1),n\in\mathbb{N}}$, $B^{(\alpha)} = (B_I^{(\alpha)})_{\ell_1+\ldots+\ell_N=n,n\in\mathbb{N}_{+0}}$ *for* $1 \le \alpha \le r$, *where*

$$\gamma_i^{(\alpha)}({b_{i1}^{(\alpha)}}^*, \cdots, {b_{iN}^{(\alpha)}}^*) = [{b_{i1}^{(\alpha)}}^*, \cdots, {b_{iN}^{(\alpha)}}^*]_Q.$$

Then there exists a regular matrix R *such that*

$$BR = \begin{pmatrix} B^{(0)} & 0 & 0 & \cdots & 0 \\ 0 & B^{(1)} & 0 & \cdots & 0 \\ & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & B^{(r)} \end{pmatrix}.$$

(Proof)
The key point of the proof is to use

$$\begin{pmatrix} \prod_{j=1}^N b_{1j}{}^{\ell_j} \\ \prod_{j=1}^N b_{2j}{}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{Hj}{}^{\ell_j} \end{pmatrix} = \begin{pmatrix} b_{11}{}^{\ell'_1} \prod_{j=2}^N b_{1j}{}^{\ell_j} & 0 & \cdots & 0 \\ 0 & b_{21}{}^{\ell'_1} \prod_{j=2}^N b_{2j}{}^{\ell_j} & \cdots & 0 \\ \vdots & \ddots & & 0 \\ 0 & 0 & \cdots & b_{H1}{}^{\ell'_1} \prod_{j=2}^N b_{Hj}{}^{\ell_j} \end{pmatrix} \begin{pmatrix} b_{11}{}^{\ell_1 - \ell'_1} \\ b_{21}{}^{\ell_1 - \ell'_1} \\ \vdots \\ b_{H1}{}^{\ell_1 - \ell'_1} \end{pmatrix},$$

and Lemma 3.

Q.E.D.

References

[1] Akaike, H.: Likelihood and Bayes procedure. Bayesian Statistics (Bernald J.M. eds.) University Press, Valencia, Spain (1980) 143–166

[2] Amari, S., Fujita, N., Shinomoto, S.: Four Types of Learning Curves. Neural Computation **4**-4 (1992) 608–618

[3] Amari, S., Murata, N.: Statistical theory of learning curves under entropic loss. Neural Computation **5** (1993) 140–153

[4] Aoyagi, M.: The zeta function of learning theory and generalization error of three layered neural perceptron. RIMS Kokyuroku, Recent Topics on Real and Complex Singularities (2006) No. 1501, pp.153-167.

[5] Aoyagi, M., Nagata, K.: Learning coefficient of generalization error of three layered neural networks and normal mixture models in Bayesian estimation (preprint).

[6] Aoyagi, M., Watanabe, S.: Resolution of Singularities and the Generalization Error with Bayesian Estimation for Layered Neural Network. IEICE Trans. J88-D-II, **10** (2005a) 2112–2124 (English version : Systems and Computers in Japan John Wiley & Sons Inc. (in press))

[7] Aoyagi, M., Watanabe, S.: Stochastic Complexities of Reduced Rank Regression in Bayesian Estimation. Neural Networks **18** (2005b) 924–933

[8] Bernstein, I. N.: The analytic continuation of generalized functions with respect to a parameter. Functional Anal. Appl., **6** (1972) 26–40

[9] Björk, J. E.: Rings of differential operators. Amsterdam: North-Holland (1979)

[10] Fukumizu, K.: A regularity condition of the information matrix of a multilayer perceptron network. Neural Networks **9**-5 (1996) 871–879

[11] Fulton, W.: Introduction to toric varieties. Annals of Mathematics Studies Princeton University Press (1993) p131

[12] Hagiwara, K., Toda, N., Usui, S.: On the problem of applying AIC to determine the structure of a layered feed-forward neural network. Proc. of IJCNN Nagoya Japan **3** (1993) 2263–2266

[13] Hartigan, J. A.: A Failure of likelihood asymptotics for normal mixtures. Proceedings of the Berkeley Conference in Honor of J.Neyman and J.Kiefer **2** (1985) 807–810

[14] Hironaka, H.: Resolution of Singularities of an algebraic variety over a field of characteristic zero. Annals of Math. **79** (1964) 109–326

[15] Kashiwara, M.: B-functions and holonomic systems. Inventions Math., **38** (1976) 33–53

[16] Kollár, J.: Singularities of pairs, Algebraic geometry-Santa Cruz 1995, Proc. Sympos. Pure Math., **62**, Amer. Math. Soc., Providence, RI, (1997 221–287

[17] Levin, E., Tishby, N., Solla, S. A.: A statistical approaches to learning and generalization in layered neural networks. Proc. of IEEE **78**-10 (1990) 1568–1674

[18] Mackay, D. J.: Bayesian interpolation. Neural Computation **4**-2 (1992) 415–447

[19] Mustata, M.: Singularities of pairs via jet schemes, J. Amer. Math. Soc. **15** (2002), 599-615.

[20] Rissanen, J.: Stochastic complexity and modeling. Annals of Statistics **14** (1986) 1080–1100

[21] Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6**-2 (1978) 461–464

[22] Sturmfels, B.: Open problems in algebraic statistics, in Emerging Applications of Algebraic Geometry, (editors M. Putinar and S. Sullivant), I.M.A. Volumes in Mathematics and its Applications, **149**, Springer, New York, (2008) 351-364

[23] Sussmann, H. J.: Uniqueness of the weights for minimal feed-forward nets with a given input-output map. Neural Networks **5** (1992) 589–593

[24] Watanabe, S.: Algebraic analysis for nonidentifiable learning machines. Neural Computation **13**-4 (2001a) 899–933

[25] Watanabe, S.: Algebraic geometrical methods for hierarchical learning machines. Neural Networks **14**-8 (2001b) 1049–1060

[26] Watanabe, S., Hagiwara, K., Akaho, S., Motomura, Y., Fukumizu, K., Okada M., Aoyagi, M.: Theory and Application of Learning System. Morikita (2005) p. 195 (Japanese)

[27] S. Watanabe, K. Yamazaki and M. Aoyagi, Kullback Information of Normal Mixture is not an Analytic Function, *Technical report of IEICE*, NC2004, 2004, 41-46.

[28] Yamanishi, K.: A decision-theoretic extension of stochastic complexity and its applications to learning. IEEE Trans. on Information Theory **44**-4 (1998) 1424–1439

[29] Yamazaki, K., Aoyagi, M., Watanabe, S.: Asymptotic Analysis of Bayesian Generalization Error with Newton Diagram, (preprint)